

**פרויקט גמר – הנדסת מחשבים**

המרכז האקדמי רופין 2024

**מגישים:**

איל רונן

גל ארז

**מנחה:**

ד"ר אריק פארן

תוכן

[מבוא 3](#_Toc181285422)

[רקע 3](#_Toc181285423)

[מוטיבציה 3](#_Toc181285424)

[מטרת הפרויקט 4](#_Toc181285425)

[סקירה ספרותית 5](#_Toc181285426)

[הפתרונות הקיימים כיום בשוק 5](#_Toc181285427)

[מקורות 6](#_Toc181285428)

[שלבי הפיתוח 7](#_Toc181285429)

[איסוף וארגון הנתונים 7](#_Toc181285430)

[תיאור המערכת לחילוץ נתונים 8](#_Toc181285431)

[תיוג ועיבוד הנתונים 8](#_Toc181285432)

[פיתוח ממשק משתמש 9](#_Toc181285433)

[שיטות וכלים 11](#_Toc181285434)

[תיאור האלגוריתמים שנבחרו 11](#_Toc181285435)

[תהליך אימון המודל ושיפור התוצאות 12](#_Toc181285436)

[Grid Search 13](#_Toc181285437)

[שימוש ב-NLP 15](#_Toc181285438)

[תיאור 15](#_Toc181285439)

[תהליך אימון המודל ושיפור התוצאות 16](#_Toc181285440)

[המשך אימון המודל 17](#_Toc181285441)

[תוצאות והערכה 18](#_Toc181285442)

[מדדים להשוואת המודלים 18](#_Toc181285443)

[תוצאות הניסויים 19](#_Toc181285444)

[דיון בתוצאות 21](#_Toc181285445)

[ממשק UI למשתמש קצה 23](#_Toc181285446)

[פונקציונאליות נוספת למשתמש הקצה 25](#_Toc181285447)

# מבוא

## רקע

בשנים האחרונות, התקדמות טכנולוגית משמעותית והתרחבות השימוש באינטרנט ובתקשורת הדיגיטלית, יצרו סביבה פורייה להתפתחות תופעות הונאה דיגיטליות, ובראשן מתקפות הפישינג. מתקפות אלו מתבצעות במגוון דרכים כמו הודעות דואר אלקטרוני, הודעות טקסט, רשתות חברתיות, ואתרי אינטרנט מזויפים. מטרתן המרכזית היא להטעות את הקורבנות ולשכנעם לחשוף מידע אישי או פיננסי רגיש, כגון סיסמאות, פרטי כרטיסי אשראי ופרטי גישה לחשבונות. התוקפים מתחזים לגורמים אמינים ומוכרים, כגון בנקים, חברות אשראי, וספקי שירותים מקוונים, ומנצלים חולשות פסיכולוגיות כמו תחושת דחיפות, פחד, ואמון בסיסי של המשתמש כדי לגרום להם לבצע פעולות אשר חושפות בפניהם מידע יקר ערך.

מתקפות הפישינג הפכו לאחד מהאיומים הנפוצים והמסוכנים ביותר בתחום אבטחת המידע. הן פוגעות לא רק במשתמשים פרטיים אלא גם בארגונים ובחברות מסחריות וגורמות לנזקים כלכליים כבדים, אובדן מידע סודי ופגיעה באמון הציבורי בשירותים דיגיטליים. התפתחות התחום מתבטאת גם בהתקדמות שיטות המתקפה: התוקפים עושים שימוש במגוון תחבולות, כולל הודעות מותאמות אישית, מניפולציות פסיכולוגיות וזיוף דפים הנראים זהים לדפים אמיתיים, כדי להערים על המשתמשים ולגרום להם לחשוף את המידע.

עם העלייה במספר המתקפות, בתחכומן ובתחכום הטכנולוגי שלהן, הצורך בכלים ובמערכות מתקדמות לזיהוי ולמניעת פישינג הפך להכרחי. מערכות המבוססות על למידת מכונה ועיבוד שפה טבעית (Natural Language Processing - NLP) מספקות כיום כלים חדשניים לזיהוי אוטומטי של ניסיונות פישינג ולמניעת גישה לא מורשית למידע רגיש. טכנולוגיות אלו מאפשרות לבצע ניתוח מעמיק של תכני ההודעות, לזהות דפוסים ותבניות מחשידים, ולהעריך את רמת הסיכון במידע המתקבל.

## מוטיבציה

המוטיבציה לפרויקט זה נובעת מהצורך הדחוף לפתח פתרונות מתקדמים ואוטומטיים שיכולים להתמודד עם האתגרים הללו, לשפר את רמת אבטחת המידע ולצמצם את הסיכון למתקפות פישינג מוצלחות. בעוד שהשיטות המסורתיות לזיהוי פישינג מתבססות על רשימות חסימה ודיווחים ידניים, שיטות אלו לרוב אינן מספקות מענה מקיף, משום שהן מתקשות להתמודד עם השינויים המתמידים במאפייני המתקפות.

המערכת שאנו מפתחים מיועדת להתמודד עם בעיות אלו באמצעות יישומים מתקדים של למידת מכונה ועיבוד שפה טבעית. המערכת תבצע ניתוח מעמיק של תכנים המתקבלים ממקורות שונים, כגון הודעות דוא"ל והודעות טקסט, ותזהה תבניות ומאפיינים ייחודיים של מתקפות פישינג. היכולות של למידת מכונה יאפשרו למערכת לזהות תבניות חשודות גם במקרים שבהם צורת המתקפה שונה ממקרים קודמים, מה שהופך את המערכת למותאמת לסביבה דינמית שבה מתקפות פישינג משתנות ומפתיעות כל הזמן.

שימוש בטכנולוגיות מתקדמות אלו לא רק משפר את דיוק הזיהוי, אלא גם מקנה יתרון משמעותי מבחינת מהירות התגובה. כאשר זמן התגובה הוא קריטי, במיוחד במתקפות אשר מטרתן היא לגנוב מידע יקר ערך, פיתוח מערכת אוטומטית שמסוגלת להגיב בזמן אמת הופך להיות הכרחי. על ידי שילוב של ניתוח שפה, תבניות טקסט ובדיקות קישורים ותכנים, המערכת שלנו מתוכננת להציע למשתמשים שכבת הגנה נוספת מול מתקפות פישינג, תוך זיהוי ונטרול של איומים בזמן אמת.

## מטרת הפרויקט

מטרת הפרויקט היא לפתח מערכת מתקדמת שתסייע בזיהוי אוטומטי של ניסיונות פישינג תוך שימוש במודלים של למידת מכונה ועיבוד שפה טבעית. המערכת תתמקד בניתוח מעמיק של תוכן ההודעות במטרה להבחין בין הודעות לגיטימיות לבין ניסיונות הונאה, תוך שהיא מעניקה למשתמשים הערכה מדויקת בנוגע לרמת הסיכון ומספקת המלצות להתנהלות בטוחה.

המערכת שאנו מפתחים נועדה להבטיח את פרטיות המידע של המשתמשים ולשפר את אבטחת המידע בסביבה הדיגיטלית המודרנית, על ידי שילוב של טכנולוגיות מתקדמות המאפשרות ניתוח מהיר ומדויק של תכנים דיגיטליים. כדי להשיג מטרה זו, ישנן מספר מטרות משנה לפרויקט:

1. **איתור וניתוח של אלמנטים מרמזים**: הפרויקט יתמקד בזיהוי וניתוח אלמנטים המרמזים על פישינג, כגון צורת הפנייה, סימנים לינגוויסטיים אופייניים להונאות ואסטרטגיות פסיכולוגיות שמתקיפים משתמשים בהן כדי להטעות את הקורבן. דגש מיוחד יינתן על ניתוח תוכן ההודעות וזיהוי קישורים חשודים, על מנת לזהות ככל האפשר את ניסיונות הפישינג.
2. **דיוק בזיהוי ניסיונות פישינג**: אחת המטרות המרכזיות היא להגיע לרמת דיוק גבוהה בזיהוי ניסיונות פישינג, תוך מזעור של תקלות והקטנת מספר ה"אזעקות שווא". לשם כך, המערכת תשתמש באלגוריתמים מתקדמים ללמידת מכונה, המתמקדים בזיהוי תבניות חוזרות בהודעות ובתכנים הייחודיים למתקפות פישינג.
3. **מהירות ניתוח ותגובה**: בכדי להבטיח הגנה מיטבית, ישנה חשיבות רבה למהירות הניתוח והתגובה של המערכת. המערכת מתוכננת לנתח הודעות בזמן אמת, לזהות ניסיונות פישינג ולהתריע עליהם מיידית, בכדי להעניק למשתמשים אפשרות להימנע מההונאה בזמן אמת.
4. **מדדים להערכת הצלחת הפרויקט**: הצלחת הפרויקט תימדד לפי מספר פרמטרים חשובים, כגון אחוזי הדיוק בזיהוי ניסיונות פישינג, מהירות התגובה של המערכת, והיכולת להתמודד עם תרחישים חדשים שלא היו ידועים מראש. מדדים אלו יעזרו לנו לשפר ולכוונן את המערכת כך שתהיה מותאמת ככל האפשר לצרכים העכשוויים ולהתפתחות האיומים בתחום אבטחת המידע.

בסיכומו של דבר, מטרת הפרויקט היא להציע פתרון שמעניק למשתמשים שכבת הגנה נוספת מול ניסיונות הונאה, תוך הגברת תחושת הביטחון בסביבה הדיגיטלית שבה הם פועלים.

# סקירה ספרותית

## הפתרונות הקיימים כיום בשוק

בשוק קיימים מספר פתרונות להתמודדות עם ניסיונות פישינג, המשלבים טכנולוגיות מתקדמות וגישות מגוונות על מנת למנוע הונאות. הפתרונות העיקריים כוללים כלים מבוססי למידת מכונה ובינה מלאכותית, עיבוד שפה טבעית, זיהוי תבניות, וטכנולוגיות נוספות. להלן סקירה של הפתרונות הקיימים:

**פתרונות מבוססי למידת מכונה ובינה מלאכותית**

* **Anti-Phishing Working Group (APWG):** ארגון בינלאומי הפועל למאבק בפשעי סייבר, המאחד גורמים מתחומים שונים, כולל תעשייה, ממשל, ואקדמיה, לצורך שיתוף פעולה בזיהוי מתקפות פישינג ובמניעתן.
* **Netcraft:** חברה המציעה שירותים נגד פישינג, כולל כלי לסרגל דפדפן המזהה אתרי פישינג, ומשתמשת במידע שנאסף לצורך יצירת מאגר אתרים מזויפים.
* **PhishLabs:** חברה המספקת שירותי ביון סייבר הכוללים איתור, ניתוח, והסרה של איומי פישינג. השירותים כוללים ניטור בזמן אמת ותגובה מהירה.
* **Microsoft's SmartScreen & Google Safe Browsing**: כלים המובנים בדפדפנים הפופולריים, המסייעים למשתמשים לזהות תוכן מזויף או מסוכן, תוך שימוש בטכנולוגיות של למידת מכונה.

**פתרונות מבוססי עיבוד שפה טבעית (NLP)**

* **Microsoft's SmartScreen & Google Safe Browsing**: שני כלים אלו משתמשים בטכנולוגיות של עיבוד שפה טבעית כדי לנתח תוכן אתרי אינטרנט ודואר אלקטרוני, לזהות תבניות לשוניות חשודות ולאתר סימנים המאפיינים הודעות פישינג.

**פתרונות מבוססי זיהוי תבניות**

* **OpenDNS**: שירות סינון רשת מבוסס ענן, המשתמש בזיהוי תבניות חוזרות על מנת לחסום אתרי פישינג ואתרים המכילים תוכן מסוכן.
* **Symantec's Anti-Phishing Toolbar**: סרגל כלים לדפדפן המשולב במערכת זיהוי תבניות, ומיועד לחסימת אתרי פישינג בזמן גלישה באינטרנט. הכלי נועד לסייע למשתמשים להימנע מאתרים מזויפים ומסוכנים.

**טכנולוגיות אחרות בשימוש**

* **AI-driven phishing detection**: פיתוח מודלים מתקדמים לניתוח תוכן דוא"ל והתנהגות משתמשים, כדי לזהות ניסיונות פישינג בזמן אמת.
* **Real-time phishing detection**: כלים וטכנולוגיות מתקדמות המסוגלות לזהות ולהגיב לניסיונות פישינג תוך שניות, במטרה להגן על משתמשים בזמן אמת.
* **Blockchain-based solutions**: פתרונות המבוססים על בלוקצ'יין ליצירת מערכות אימייל מבוזרות ובטוחות, המקטינים את הסיכון להונאות מסוג פישינג באמצעות אימות מקור המידע ומעקב בלתי ניתן לשינוי.

## מקורות

**1. Anti-Phishing Training Effectiveness**:

* **Source**: ["Don’t click: towards an effective anti-phishing training"](https://hcis-journal.springeropen.com/articles/10.1186/s13673-020-00237-7) from Human-centric Computing and Information Sciences, SpringerOpen.
* **Summary**: This article (SpringerOpen, 2020) delves into the effectiveness of anti-phishing training programs. It underscores the necessity of educating users about phishing tactics and offers a comparative analysis of various training methods, emphasizing their role in mitigating phishing risks.

**2. Comprehensive Survey of Phishing Attacks**:

* **Source**: ["A Comprehensive Survey of Phishing Attacks: Characteristics, Detection, and Prevention"](https://www.mdpi.com/2227-7080/11/6/161) from MDPI.
* **Summary**: Providing a thorough survey on phishing attacks (MDPI, 2023), this paper outlines the key characteristics of phishing, detection techniques, and preventive measures. It’s a valuable resource for understanding the intricate details of phishing strategies and the methodologies employed in their detection and prevention.

**3. Current Trends in Phishing**:

* **Source**: [Cybersecurity Magazine.](https://cybersecurity-magazine.com/?s=phishing)
* **Summary**: Cybersecurity Magazine offers a wealth of articles and reports on contemporary phishing trends. It covers the use of IoT devices in phishing, the emergence of ransomware, and deepfake technology, providing a nuanced understanding of modern phishing strategies and their technological evolution.

**4. Integration of NLP in Cybersecurity**:

* **Source**: [The Cyber Express.](https://thecyberexpress.com/?s=phishing)
* **Summary**: The integration of Natural Language Processing (NLP) in cybersecurity, as discussed in The Cyber Express, illustrates the potential of NLP in enhancing threat communication and response. This perspective is crucial for understanding how advanced technologies like AI and NLP can be leveraged in the fight against phishing.

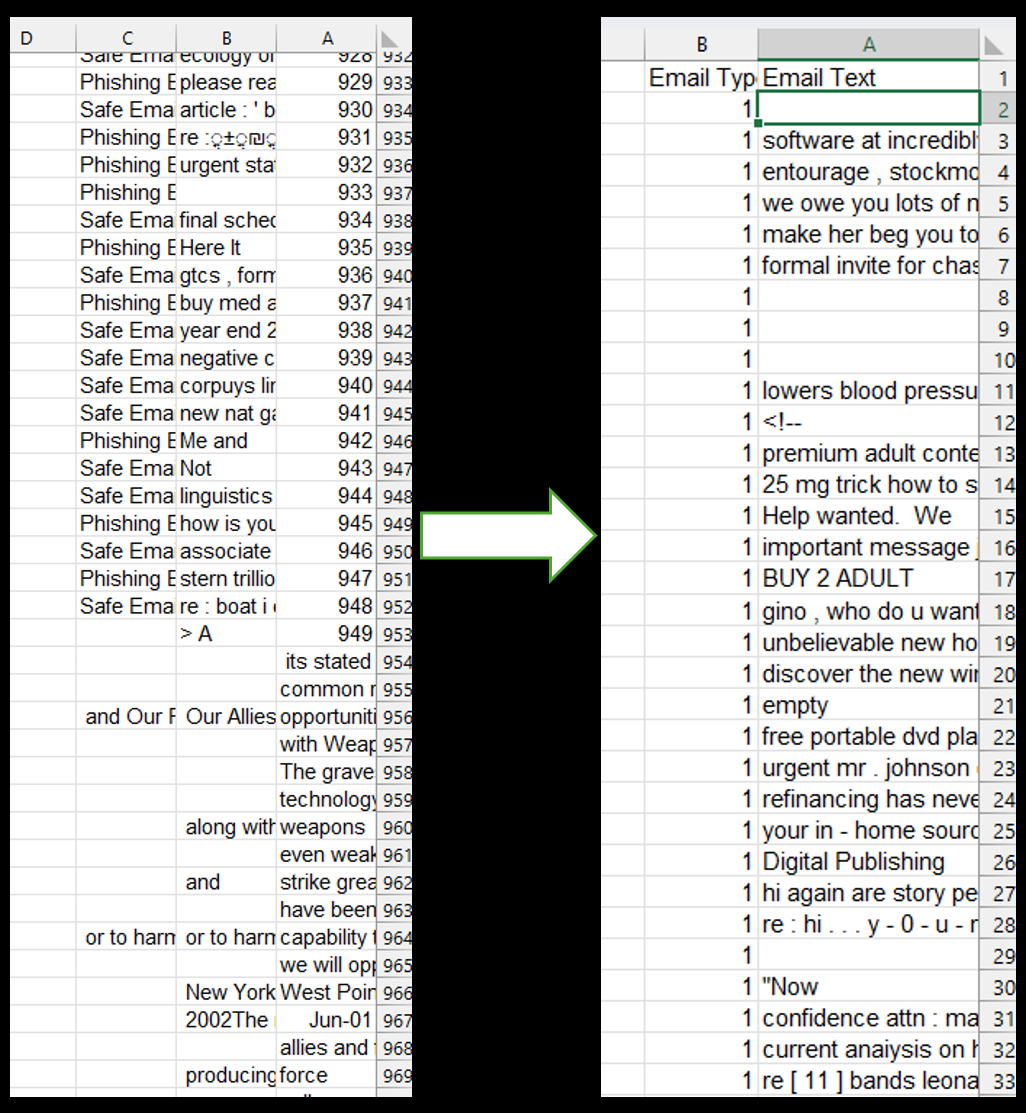
# שלבי הפיתוח

## איסוף וארגון הנתונים

השלב הראשוני בפרויקט PhishGuard AI התמקד באיסוף מקורות נתונים רלוונטיים לאימון המודלים, אשר ישמשו לזיהוי הודעות פישינג. כחלק מתהליך זה, ערכנו חיפוש במאגרים שונים הזמינים באינטרנט, עד שלבסוף בחרנו להשתמש במאגר הנתונים של פלטפורמת Hugging Face שהינה פלטפורמה שיתופית המוקדשת לתחום למידת המכונה. בחירה זו נבעה מכמה סיבות עיקריות: ראשית, המאגר היה יחסית נגיש וקל לעיבוד; שנית, ההודעות במאגר נאספו מדאטה אמיתי שהצטבר לאורך זמן, מה שהבטיח את הרלוונטיות והגיוון של הנתונים שנבחרו. בסופו של דבר, הנתונים שנבחרו כללו 18,000 הודעות דואר אלקטרוני, כאשר כ-8,000 מתוכן סווגו כהודעות פישינג והשאר כהודעות בטוחות.

עם זאת, הנתונים שנאספו לא היו מאורגנים בצורה שתאפשר שימוש ישיר במודלים של למידת מכונה. לכן, היה צורך בביצוע תהליך ראשוני של ניקוי וארגון נתונים כדי להתאימם לשימוש יעיל ואפקטיבי. תהליך ניקוי וארגון הנתונים בוצע בסביבת העבודה Jupyter Notebook תוך שימוש בספריית Pandas הנפוצה בתחום ניתוח הנתונים בשפת Python. תחילה, ייבאנו את הנתונים לקובץ DataFrame והסרנו ערכים חסרים שהיו עלולים לפגוע באיכות המודלים ובדיוק התחזיות שלהם. בנוסף, הסרנו רווחים מיותרים וביצענו המרה של סיווג ההודעות למספרים (“Phishing email" ל-1 ו-"Safe email" ל-0), על מנת להבטיח מבנה קונסיסטנטי ומתאים לעיבוד במודלי למידת מכונה.

תהליך זה היה שלב קריטי בפרויקט, מאחר והוא הבטיח את האיכות והאמינות של הנתונים ובכך השפיע ישירות על הצלחת המודלים שהוכנו. בסיום תהליך הניקוי, המידע נשמר בקובץ CSV כדי להבטיח גישה קלה ונוחה לשימוש במהלך אימון המודלים. כך הוכנה הקרקע לאימון יעיל של המודלים וליכולת ההבחנה בין הודעות פישינג לבין הודעות בטוחות באופן אפקטיבי.



## תיאור המערכת לחילוץ נתונים

השלב הבא בפרויקט כלל פיתוח מערכת לחילוץ נתונים מתוך הודעות דואר אלקטרוני, שמטרתה להעשיר את תהליך עיבוד הנתונים ולספק מידע נרחב ומעמיק לצורך אימון המודלים לזיהוי הודעות פישינג. מערכת זו נועדה להפיק מאפיינים ייחודיים ונתונים סטטיסטיים מתוך הטקסט של ההודעות, במטרה לשפר את יכולות הזיהוי וליצור תהליך זיהוי יעיל ומדויק יותר להודעות פישינג בשלב האימון של המודלים.

לשם פיתוח המערכת נעשה שימוש במגוון ספריות ותוכנות מתקדמות לעיבוד טקסט, כגון NLTK וספריות נוספות בתחום הבינה המלאכותית, אשר מאפשרות לנתח את תוכן ההודעות ברמות שונות. באמצעות כלים אלו ביצענו ניתוח מעמיק של הטקסט, שכלל ניתוח תחבירי, זיהוי ישויות ותבניות לשוניות חוזרות, הערכת רמת הקריאות של הטקסט וניתוח הסנטימנט המובע בו.

הצורך בפיתוח מערכת זו נבע מכך שהנתונים הגולמיים שנאספו לא היו מסודרים ומאורגנים באופן שמתאים ישירות לשימוש באימון המודלים של למידת המכונה. לכן, היה צורך בתהליך של חילוץ מאפיינים מתוך ההודעות, שיאפשר הבנה מעמיקה יותר של מבנה ההודעות ותוכנן ויסייע בזיהוי תבניות המאפיינות הודעות פישינג לעומת הודעות לגיטימיות. המידע המופק באמצעות המערכת מיועד להעשיר את תהליך האימון, להוסיף ערך להודעות טקסט בסיסיות, ולספק תובנות מעמיקות יותר שניתן לשלב במסגרת תהליך הזיהוי.

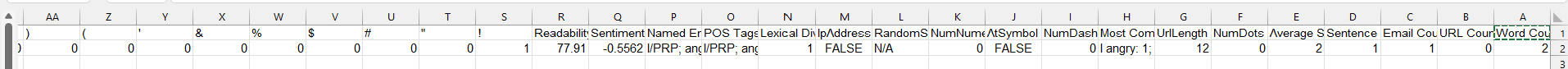
## תיוג ועיבוד הנתונים

בתהליך של תיוג ועיבוד הנתונים שהופקו מתוך הודעות הדואר האלקטרוני. המטרה הייתה להפיק מאפיינים (Features) שיסייעו לאמן את המודלים באופן מדויק ויעיל יותר, ולשפר את יכולת הזיהוי של הודעות פישינג.

להלן הפרמטרים שנבחרו לחילוץ, יחד עם הסבר קצר על תרומתם הפוטנציאלית למודל:

1. **כמות המילים :(Word Count)** כמות המילים בהודעה יכולה לשקף את אורך ההודעה, מאחר והודעות פישינג רבות הן קצרות וממוקדות, במטרה לגרום לנמען לפעול במהירות.
2. **כמות כתובות URL (URL Count):** הודעות פישינג נוטות לכלול קישורים מזויפים שנראים אמינים. זיהוי מספר כתובות URL בהודעה עשוי להצביע על חשד לניסיון פישינג.
3. **כמות כתובות דואר אלקטרוני (Email Count):** ריבוי כתובות דוא"ל בהודעה עלול להוות סימן לפישינג, מכיוון שתוקפים מנסים לעתים לשוות להודעה מראה אמין יותר באמצעות שימוש במספר כתובות.
4. **כמות המשפטים (Sentence Count) :** כמות המשפטים מאפשרת הבנה של מבנה הטקסט. הודעות פישינג עשויות להכיל משפטים קצרים ותכליתיים, במטרה לייצר תחושת דחיפות אצל הקורבן.
5. **אורך ממוצע של משפטים (Average Sentence Length) :** אורך ממוצע של משפטים עשוי לשמש לזיהוי הסגנון הלשוני של ההודעה, כאשר הודעות פישינג לעיתים כוללות משפטים פשוטים וקצרים.
6. **גיוון לשוני (Lexical Diversity)** **:** מדד לגיוון הלשוני של ההודעה, המייצג את היחס בין מספר המילים הייחודיות לבין מספר המילים הכולל בטקסט. הודעות פישינג עשויות להשתמש במילים חוזרות על מנת לייצר תחושת אמינות.
7. **תדירות בי-גרמים (Most Common Bigrams)** **:** זיהוי הצירופים הנפוצים ביותר של שתי מילים (בי-גרמים) עשוי להצביע על תבניות לשוניות שמאפיינות הודעות פישינג.
8. **ניתוח תחבירי (POS Tagging)** **:** זיהוי חלקי הדיבר של כל מילה בטקסט (כגון פועל, שם עצם). הודעות פישינג עשויות לכלול מבנה תחבירי לא רגיל או לא תקין אשר ניתן לזהותו באמצעות ניתוח זה.
9. **זיהוי ישויות (Named Entity Recognition)** **:** זיהוי שמות, כתובות, מוסדות וכדומה. לעיתים הודעות פישינג כוללות שמות של חברות או גורמים מפורסמים בניסיון להטעות את הקורבן.
10. **ניתוח סנטימנט (Sentiment Analysis)** **:** זיהוי הטון הרגשי של ההודעה (חיובי, שלילי, נייטרלי). הודעות פישינג עשויות להשתמש בטון חיובי מופרז או בטון מאיים כדי לגרום לנמען לפעול.
11. **ציון קריאות (Readability Score)** **:** הערכת רמת הקריאות של ההודעה באמצעות נוסחת קריאות של פלש (Flesch Reading Ease) . הודעות פישינג נוטות להיות פשוטות להבנה על מנת לפנות למספר רחב של אנשים.
12. **מספר תווים מספריים (NumNumericChars)** **:** מספר הספרות בהודעה. הודעות פישינג רבות כוללות מספרים כחלק מניסיון להטעות, כמו מספרי חשבון או מזהי עסקאות.
13. **זיהוי כתובות IP (Identify IP Address):** זיהוי אם קיימת כתובת IP בטקסט. כתובות IP עשויות להופיע בהודעות פישינג בניסיון להטעות את המשתמש ולשוות להודעה מראה טכני ואמין.
14. **נוכחות תווים מיוחדים, כגון '@' ו'-'** : זיהוי תווים מיוחדים מסוימים כגון '@' עשוי לסייע בזיהוי מיילים מזויפים הכוללים התחזות, או בהודעות שבהן נעשה שימוש בריבוי סימנים כדי להטעות.

תהליך זה של חילוץ ותיוג מאפיינים מהווה שלב קריטי לצורך אימון המודלים, שכן הוא מאפשר להם לזהות תבניות חוזרות שמאפיינות הודעות פישינג ולהבחין בינן לבין הודעות לגיטימיות.



## פיתוח ממשק משתמש

על מנת לחלץ את המידע בצורה נוחה פיתחנו ממשק משתמש גרפי (GUI) שמטרתו להקל על המשתמש בביצוע ניתוחי הודעות דואר. הממשק פותח באמצעות ספריית Tkinter בשפת Python ומציע מספר פיצ'רים חשובים שמאפשרים לבצע ניתוח מתקדם של הודעות טקסט.

הפיצ'רים המרכזיים של ממשק המשתמש כוללים:

1. **הזנת טקסט לניתוח** - הממשק מאפשר הזנת טקסט חופשי של הודעת דוא"ל עבור ניתוח, באמצעות תיבת טקסט גלילה נוחה.
2. **הזנת כתובת דוא"ל או כתובת URL** -שדה המיועד להזנת כתובת URL או כתובת דוא"ל לצורך ניתוח מרכיבי ההודעה, וזיהוי קישורים חשודים.
3. **טעינת קבצים להודעה** - הממשק כולל אפשרות לטעינת קבצים בפורמט .txt או .msg המאפשרים למשתמש לבצע ניתוח של הודעות המגיעות ממקור חיצוני.
4. **כפתור "בצע ניתוח"** - הממשק כולל כפתור לניתוח ההודעה המוזנת, אשר מבצע חילוץ נתונים ומתן הערכה מיידית בנוגע למאפייני ההודעה.
5. **שמירת תוצאות לקובץ CSV** -תכונה זו מאפשרת למשתמש לשמור את תוצאות הניתוח כקובץ CSV לצורך ניתוח נוסף או שמירת רישום של הניתוחים שבוצעו.

הממשק נועד להיות פשוט וקל לשימוש, תוך שמירה על תצוגה אינטואיטיבית של כל המידע המופק מההודעה. המשתמש מקבל מידע על כמות המילים, מספר הקישורים, ניתוח הסנטימנט, רמת הקריאות של ההודעה, ועוד נתונים חשובים אחרים שמסייעים בזיהוי הודעות פישינג.

תמונה שמכילה טקסט, תוכנה, דף אינטרנט, אתר

התיאור נוצר באופן אוטומטי

ממשק לחילוץ מידע מההודעות

תמונה שמכילה טקסט, צילום מסך, קו, מספר

התיאור נוצר באופן אוטומטי

תוצאות הניתוח של הדאטה

# שיטות וכלים

## תיאור האלגוריתמים שנבחרו

**רגרסיה לוגיסטית (Logistic Regression)**

* **למה נבחר:** רגרסיה לוגיסטית היא אלגוריתם פשוט ויעיל המאפשר לבצע סיווג בינארי. האלגוריתם מבצע ניבוי על פי משקל שמחושב לכל אחת מהתכונות, ומתאים במיוחד כאשר יש קשר ליניארי בין התכונות.
* **מאפיינים:** מודל זה מבוסס על הפונקציה הלוגיסטית, המספקת סיכוי לניבוי של מחלקה מסוימת (פישינג או הודעה בטוחה). אחד היתרונות הגדולים של רגרסיה לוגיסטית הוא האפשרות להסביר ולפרש את ההחלטות שהתקבלו על ידי המודל, שכן הוא מספק משקל ברור לכל תכונה.

**KNN (k-Nearest Neighbors)**

* **למה נבחר:** נבחר מאחר שהוא נחשב למודל גמיש ומתאים במיוחד לסיווג כאשר אין הנחה על התפלגות הנתונים. מדובר במודל פשוט להבנה ולביצוע.
* **מאפיינים:** KNN הוא אלגוריתם שאינו דורש שלב למידה משמעותי – כל תהליך הסיווג מתבצע בשלב הניבוי. האלגוריתם מתבצע על ידי חיפוש ה"קירובים" הקרובים ביותר לנקודה החדשה ונתינת המחלקה על סמך רוב הקירובים. חיסרון מרכזי הוא הצורך בשמירת כל הנתונים לצורך הניבוי, מה שעשוי לגרום לעומס חישובי.

**SVM (Support Vector Machine)**

* **למה נבחר**: SVM ידוע כאלגוריתם יעיל במיוחד במצבים שבהם יש צורך להפריד בין שתי מחלקות בצורה חדה וברורה, גם כאשר המידע המתקבל מורכב ולא ליניארי.
* **מאפיינים:** האלגוריתם בונה היפר-מישורים במרחב המאפיינים, המפרידים בצורה הטובה ביותר בין המחלקות השונות. השימוש בגרעין (kernel) מאפשר לו להתמודד עם נתונים לא ליניאריים ובכך מאפשר סיווג טוב יותר במקרים מורכבים.

**Random Forest**

* **למה נבחר**: Random Forest נבחר בשל יכולתו להתמודד עם תכונות מרובות ולסווג בצורה מדויקת גם כאשר קיימות אינטראקציות מורכבות ביניהן.
* **מאפיינים**: מדובר באנסמבל של עצי החלטה אשר מבוצעים במקביל ולאחר מכן מתבצע שילוב של התוצאות שלהם, מה שמאפשר יציבות גבוהה ועמידות בפני נתונים רעשיים. כל עץ ביער מתבצע על חלק שונה של הנתונים, מה שמונע תופעות של התאמת יתר (overfitting).

**Gradient Boosting**

* **למה נבחר:** Gradient Boosting נבחר בשל יכולתו לזהות דפוסים מורכבים תוך כדי תיקון טעויות שבוצעו במודלים קודמים. הוא נחשב לאחד האלגוריתמים המובילים בסיווג נתונים.
* **מאפיינים**: מודל זה מבוסס על סדרת עצי החלטה, כאשר כל עץ חדש מנסה לתקן את השגיאות של העצים הקודמים. התוצאה היא מודל חזק המסוגל לספק תוצאות מדויקות בסיווג הודעות פישינג, אך דורש יותר זמן חישוב בהשוואה למודלים אחרים.

**XGBoost**

* **למה נבחר**: XGBoost הוא שיפור על Gradient Boosting ונבחר בשל יכולתו לעבוד עם כמויות גדולות של נתונים ובמהירות גבוהה.
* **מאפיינים**: זהו אלגוריתם Boosting משופר עם אופטימיזציות רבות שמאפשרות עיבוד מהיר ויעיל, כמו גם טיפול טוב יותר בתכונות ריקות. XGBoost נחשב לאחד האלגוריתמים החזקים ביותר בתחרויות למידת מכונה (Kaggle) , ומתאים היטב לפרויקטים שבהם יש צורך בדיוק גבוה וביצועים טובים.

**נאיב בייס (Naive Bayes)**

* **למה נבחר**: נאיב בייס נבחר בשל מהירותו והפשטות היחסית שלו, כמו גם יכולתו לטפל בטקסטים. מתאים במיוחד לניתוח תכנים טקסטואליים כמו הודעות אימייל.
* **מאפיינים**: מודל זה מתבסס על ההנחה שכל תכונה בלתי תלויה בתכונה אחרת (הנחת נאיביות). למרות ההנחה הזו, נאיב בייס מצליח לספק ביצועים טובים במקרים רבים, במיוחד כאשר מדובר בטקסטים.

**עץ החלטה (Decision Tree)**

* **למה נבחר**: עץ החלטה נבחר מאחר שהוא פשוט להבנה ולהסבר ומהווה מודל בסיסי לביצוע סיווג.
* **מאפיינים**: עץ החלטה בנוי כסדרה של שאלות ותשובות בינאריות, אשר כל אחת מהן מובילה לסיווג הסופי. הפשטות של עץ החלטה מאפשרת לפרש את תהליך קבלת ההחלטות בקלות, אך חסרונו העיקרי הוא נטייה להטות יתר במקרה של עצים עמוקים מדי.

**רשת נוירונים מלאכותית (Neural Network)**

* **למה נבחר**: רשת נוירונים נבחרה בשל היכולת שלה לזהות דפוסים מורכבים מאוד ולא ליניאריים, ולכן מתאימה לזיהוי דפוסים מורכבים הקיימים בהודעות פישינג.
* **מאפיינים**: רשת נוירונים מלאכותית מחקה את פעולתו של מוח אנושי, ומתאימה למגוון רחב של סוגי בעיות, כולל זיהוי דפוסים מורכבים בטקסטים. רשתות הנוירונים מתבססות על שכבות של נוירונים אשר מאפשרות גילוי תכונות מורכבות ובלתי נראות למודלים אחרים.

### תהליך אימון המודל ושיפור התוצאות

תהליך אימון המודלים בפרויקט ושיפור התוצאות כלל שלבים רבים, שכל אחד מהם נועד להבטיח שהמודלים יהיו מדויקים, יעילים ועמידים ככל האפשר בזיהוי הודעות פישינג.

בשלב הראשוני של אימון המודל, נעשה שימוש במגוון אלגוריתמים של סיווג, כפי שתואר קודם לכן, ובוצעה חלוקת הנתונים לסטים של אימון ובדיקה. כדי למקסם את ביצועי המודלים ולהשיג את התוצאות המדויקות ביותר, נעשה שימוש בגישת "Grid Search" לאופטימיזציה של הפרמטרים.

### Grid Search

Grid Search היא מתודולוגיה מרכזית בתחום למידת המכונה, המשמשת למציאה אוטומטית של ההגדרות המיטביות עבור מודלים שונים. במסגרת הפרויקט, יישמנו את שיטת ה-Grid Search כדי לבחון באופן שיטתי ומקיף מגוון רחב של שילובי פרמטרים עבור כל אחד מהמודלים שנבחנו.

**פרמטרים עיקריים שנבחנו**

**מספר העצים במודלים מבוססי Ensemble (n\_estimators)**

פרמטר זה, הרלוונטי במיוחד למודלים כגון Random Forest ו-XGBoost, קובע את מספר עצי ההחלטה במודל. בעוד שהגדלת מספר העצים עשויה לשפר את דיוק המודל, נדרש איזון מול זמן החישוב וניצול המשאבים.

**עומק מקסימלי של העץ (max\_depth)**

פרמטר חשוב המשפיע על מורכבות המודל. עומק נמוך מדי עלול לגרום למודל להיות פשטני מדי (underfitting), בעוד עומק גבוה מדי עלול לגרום למודל ללמוד את הנתונים בצורה מדויקת מדי (overfitting). באמצעות Grid Search מצאנו את העומק האידיאלי שמאזן בין שתי אפשרויות אלו.

**קצב למידה (learning\_rate)**

במודלים כמו Gradient Boosting ו-XGBoost, פרמטר זה קובע עד כמה מהר המודל לומד מהנתונים, ומשפיע על היכולת שלו להגיע לתוצאות מדויקות.

**פרמטר בקרה (C)**

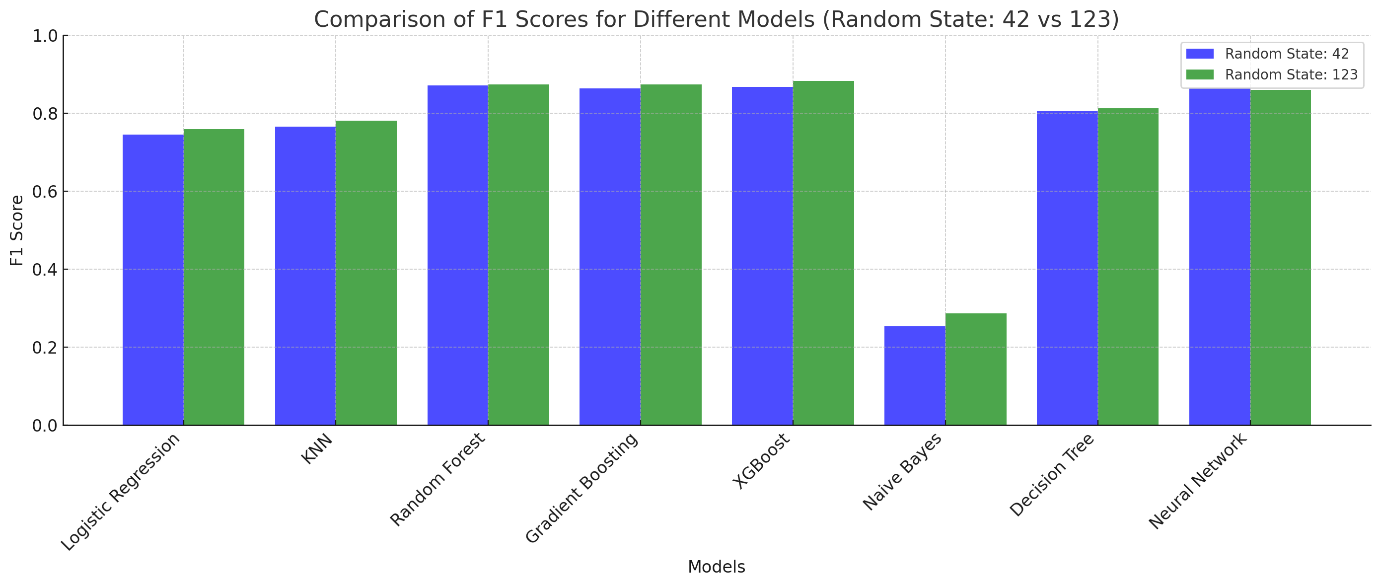
במודל SVM, פרמטר C קובע את מידת הדגש שניתן לטעויות סיווג. ערך גבוה יותר נותן חשיבות רבה יותר לדיוק בסיווג, בעוד ערך נמוך יותר מאפשר למודל להיות גמיש יותר.

**מבנה הרשת (hidden\_layer\_sizes)**

ברשתות נוירונים, פרמטר זה קובע את המבנה הפנימי של הרשת ומספר היחידות החישוביות בכל שכבה, ובכך משפיע על יכולת המודל ללמוד דפוסים מורכבים בנתונים.

**בדיקת יציבות המודלים**

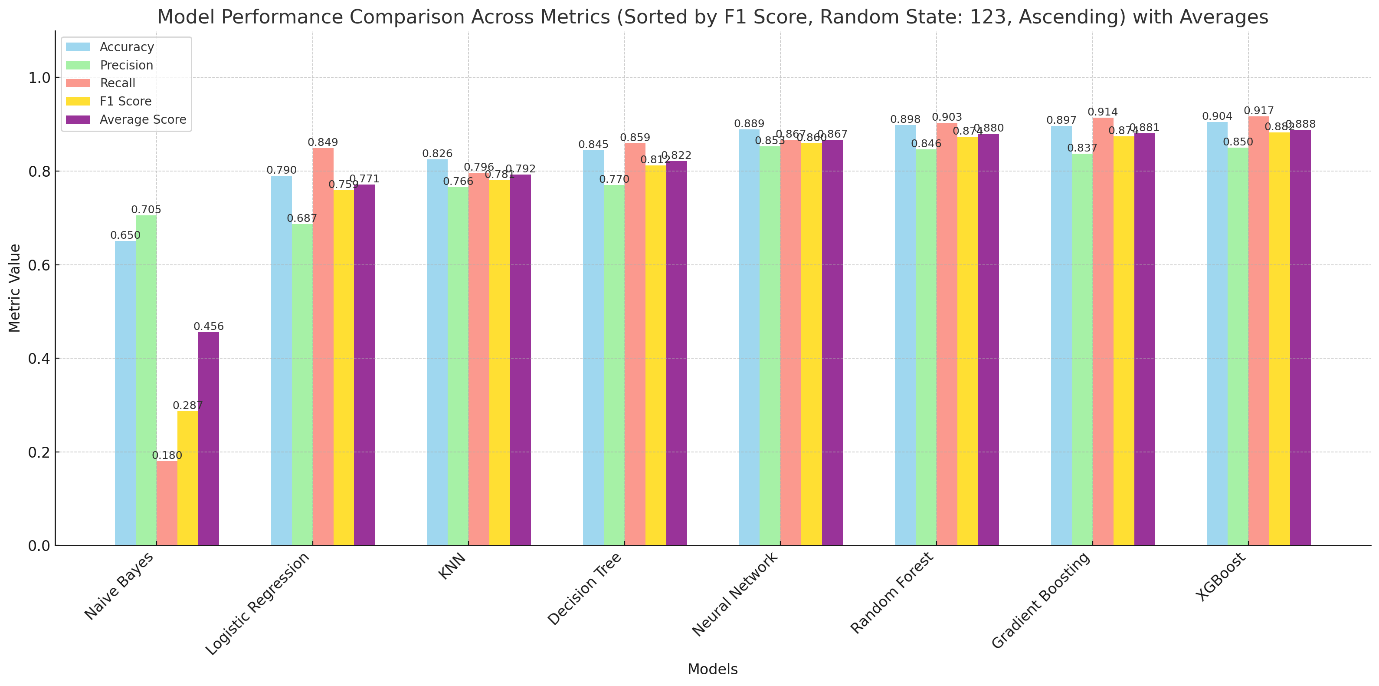
כחלק מתהליך האימון והבדיקה, בחנו את ביצועי המודלים באמצעות מספר ערכי Random State (42 ו-123). גישה זו נועדה לוודא שהמודלים פועלים באופן יציב ואמין גם כאשר הנתונים מחולקים בצורות שונות לקבוצות אימון ובדיקה. תהליך זה מחזק את האמינות של המודלים שפותחו במסגרת הפרויקט ומבטיח שהתוצאות אינן מוטות בגלל חלוקה מסוימת של הנתונים.



**תוצאות ושיפור**

לאחר תהליך Grid Search וביצוע האופטימיזציה של הפרמטרים, מדדנו את ביצועי המודלים בעזרת מדדים שונים כמו (Accuracy), (Precision), (Recall) ו-(F1 Score) . מדדים אלו אפשרו לנו להבין את הביצועים של כל מודל ולזהות את המודל המוביל בזיהוי ניסיונות הפישינג.

לבסוף, המודל שנמצא כמתאים ביותר לפרויקט היה  **XGBoost** , אשר הציג ביצועים גבוהים במיוחד בכל המדדים שנבדקו. השימוש בטכניקת ה-Grid Search ובחינה של ערכי Random State שונים הבטיח שהמודל יעבוד בצורה מיטבית ויעילה על מגוון רחב של הודעות דואר אלקטרוני, גם במצבים חדשים ולא צפויים.



# שימוש ב-NLP

## תיאור

בפרויקט בחרנו לבדוק גם גישה מבוססת על עיבוד שפה טבעית (NLP) במקביל לשיטות אחרות של למידת מכונה. ההבדל המרכזי בין הגישה של למידת מכונה לבין גישת ה-NLP טמון בשיטות בהן נעשה עיבוד של הטקסטים. בגישה של למידת מכונה אנו מנסים למצוא תבניות מסוימות באמצעות מאפיינים שונים מתוך הנתונים, בעוד שגישת ה-NLP שמה דגש על הבנה עמוקה של השפה והמשמעות שמאחורי המילים בהקשרן. המטרה היא להפיק תובנות מעמיקות יותר, שמאפשרות לזהות גם דקויות בשפה שעשויות להעיד על ניסיון פישינג.

**מודל DistilBERT**

בפרויקט השתמשנו במודל NLP מתקדם מסוג DistilBERT , המהווה גרסה קלה ויעילה של מודל BERT. DistilBERT פותח במטרה לאפשר שימוש במודלים טרנספורמטיביים באופן מהיר יותר, ללא פגיעה משמעותית בביצועים. מודלים טרנספורמטיביים מסוגלים להבין הקשרים בין מילים בטקסטים בצורה טובה יותר, מה שמאפשר לזהות ניסיונות פישינג המתחזים לשפה רגילה.

**שלבי הטמעת המודל**

**הכנת הנתונים**

השלב הראשון היה להגדיר ולהכין את הנתונים. הטקסטים של הודעות הדוא"ל חולקו לשני חלקים - אחד לאימון המודל והשני לבדיקה (Validation) . חשוב להדגיש כי במאגר הנתונים שלנו כמות הודעות הפישינג הייתה יחסית גדולה בהשוואה למציאות, שבה מעט מאוד הודעות הן הודעות פישינג. כדי להתמודד עם חוסר ההתאמה בין הדאטה לבין השכיחות האמיתית, ביצענו התאמות שמביאות בחשבון את השכיחות הנמוכה של הודעות פישינג במציאות. השתמשנו בטכניקת דגימה מחדש (Undersampling) באמצעות RandomUnderSampler, שבמסגרתה הקטנו את כמות ההודעות הבטוחות כדי ליצור מאגר מאוזן יותר לאימון המודל. פעולה זו נועדה להבטיח שהמודל ילמד בצורה נכונה ומדויקת את מאפייני הודעות הפישינג ולא יהיה מוטה לעבר הסיווג של הודעות בטוחות בלבד, דבר אשר תורם ליכולת המודל להתמודד עם המצב האמיתי בצורה מיטבית.

**טוקניזציה (Tokenization)**

טוקניזציה היא תהליך שמטרתו לחלק את הטקסט לחלקים קטנים יותר שהמודל יכול להבין ולעבד בקלות. למשל, מילים או ביטויים חשובים. בשלב זה השתמשנו בטוקנייזר של DistilBERT ,אשר הפך את הטקסט לאוסף של מילים (טוקנים), וכן הוסיף טוקנים מיוחדים המאפשרים למודל להבין את ההקשר הכללי שבו הטקסט מופיע.

**בניית ה-Dataset וטעינת הנתונים**

כדי לייעל את תהליך האימון, היה עלינו לבנות מערך נתונים (Dataset) מיוחד המכיל את כל ההודעות והסיווג שלהן (האם מדובר בהודעת פישינג או הודעה בטוחה). המידע הועבר למבנה בו ניתן להשתמש במסגרת האימון, כך שכל הודעה עברה תהליך של המרה, חלוקה, והכנה בצורה שתאפשר למודל ללמוד טוב יותר את המאפיינים שלה.

**אימון המודל**

תהליך האימון של המודל התבצע על גבי יחידת עיבוד גרפית (GPU). המטרה הייתה ללמד את המודל להבחין בין הודעות פישינג לבין הודעות רגילות על בסיס הדפוסים שהוא מזהה בטקסט. כדי להבטיח שהתהליך יהיה יציב, נעשה שימוש בשיטות שונות המאפשרות התאמה טובה של קצב הלמידה ושמירה על היכולת של המודל להתאים עצמו בצורה הדרגתית לנתונים.

**הערכת ביצועי המודל**

לאחר האימון, ביצענו הערכה של ביצועי המודל על בסיס סט הנתונים ששמרנו לבדיקה. ההערכה כללה שימוש במדדים שונים, כמו דיוק, דיוק חיובי (Precision) ,יכולת גילוי (Recall) , ומדד F1 המשלב את השניים. מדדים אלה נתנו לנו אינדיקציה לגבי יכולת המודל לזהות בצורה מדויקת ניסיונות פישינג מבלי לייצר יותר מדי זיהויים שגויים.

**שמירת המודל והערכתו לשימוש עתידי**

כדי לאפשר שימוש עתידי במודל המאומן, שמרנו אותו בפורמט מתאים ובנינו קובץ תצורה המפרט את כל ההגדרות בהן השתמשנו לאימון המודל. זה יאפשר לנו להפעיל את המודל גם בעתיד לצורך ניתוח הודעות נוספות.

**יתרונות השימוש ב-NLP ומודלים טרנספורמטיביים**

שימוש בעיבוד שפה טבעית ומודלים טרנספורמטיביים כמו DistilBERT מאפשר להבין הקשרים בטקסטים המופיעים בהודעות פישינג. מודלים מסוג זה מבוססים על ארכיטקטורת טרנספורמציה שמסוגלת "לזכור" חלקים מהטקסט ולהבין את ההקשרים ביניהם. זהו יתרון משמעותי כאשר אנו מנסים להבחין בין טקסט רגיל לבין טקסט מתחזה שנועד להונות.

גישה זו מאפשרת לנו לבצע הבנה עמוקה יותר של השפה ושל הסימנים המעידים על פישינג, לזהות מילים בעייתיות או יוצאות דופן, לנתח את הכוונות שבטקסט, ולהפיק תובנות נוספות שיעזרו בזיהוי יעיל יותר של ניסיונות הונאה. כל זאת, תוך שימוש בטכנולוגיות מתקדמות שמאפשרות להפחית את כמות השגיאות ולשפר את הבטיחות של משתמשי הדוא"ל.

### תהליך אימון המודל ושיפור התוצאות

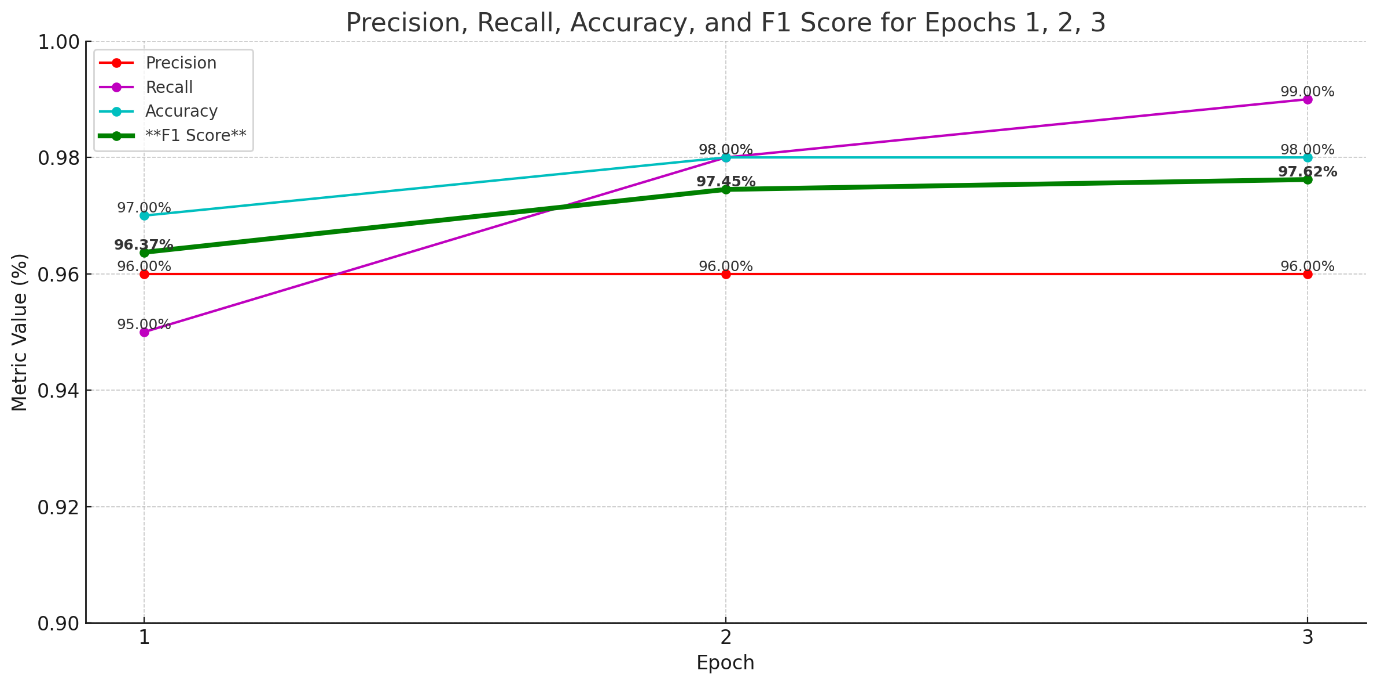
תהליך אימון המודל כלל מספר שלבים שמטרתם להכשיר את מודל DistilBERT כדי לזהות הודעות פישינג בצורה מדויקת ויעילה. האימון התבצע במתודולוגיה מבוקרת באמצעות חלוקת הנתונים לשני חלקים: אחד לצורך אימון המודל והשני לצורך בדיקה ואימות. חלוקה זו מאפשרת לוודא שהמודל אינו "זוכר" את הנתונים שעליהם אומן, אלא שהוא מסוגל להכליל את הידע שלו לנתונים חדשים שמעולם לא ראה.

תהליך האימון התבצע לאורך מספר מחזורים הנקראים **Epochs**.כל Epoch מייצג מחזור שלם שבו המודל נחשף לכל סט הנתונים פעם אחת. בשלב זה, המודל לומד לזהות דפוסים ומאפיינים של הודעות פישינג והודעות לגיטימיות. מספר ה-Epoches שהשתמשנו בו בפרויקט היה 3, וזאת כדי לאפשר למודל ללמוד את מאפייני הנתונים מבלי להגיע למצב של "התאמת יתר" (Overfitting), שבו המודל מתאים את עצמו יתר על המידה לנתוני האימון, וכך נפגעת יכולתו להכליל על נתונים חדשים.

במהלך כל Epoch המודל מתקן את המשקלים (weights) שלו באמצעות אופטימייזר (AdamW). במקרה זה, על בסיס הטעויות שהוא מבצע בניבוי הסיווג של הנתונים. לאחר כל שלב של אימון בוצעה הערכת ביצועים של המודל על סט האימות. הערכה זו כוללת מדידת מדדים כמו דיוק (Accuracy), Precision, Recall, ומדד F1. מטרת הערכה זו היא לוודא שהמודל אכן מצליח לזהות נכון את הודעות הפישינג וגם להימנע מטעויות של סיווג שגוי.

התוצאות שהתקבלו במהלך האימון הראו שיפור משמעותי בדיוק בזיהוי הודעות פישינג, כשהמדדים השתפרו עם כל מחזור אימון נוסף. לדוגמה מדד ה-F1 המשלב בין Precision ו-Recall, הראה מגמה של עלייה עם התקדמות האימון, מה שמעיד על כך שהמודל הצליח ללמוד כיצד להבחין בין הודעות פישינג להודעות לגיטימיות בצורה מדויקת ויציבה יותר.

בסיום תהליך האימון, המודל המאומן נשמר לצורך שימוש עתידי, וכן נשמרו גם קובצי התצורה שלו (כגון config.json) כדי לאפשר טעינה מחדש של המודל לצורך חיזוי עתידי. התוצאה שהתקבלה היא מודל המסוגל לסווג הודעות דוא"ל כהודעות פישינג או כהודעות בטוחות בצורה אוטומטית, ובכך להגביר את האבטחה עבור המשתמשים.



## המשך אימון המודל

בהתייחס לסוגיית המשך אימון מודל ה DistilBERT-ולחשש הנובע מתופעת ה overfitting-, חשוב להדגיש את ההיבטים המתודולוגיים והאמפיריים של תהליך האימון הנוסף. האימון התבצע על מאגר נתונים ייחודי המורכב מהודעות דואר אלקטרוני אותנטיות, אשר נאספו מתוך תקשורת יומיומית של משתמשים בסביבה מבצעית אמיתית. בעוד שניתן לטעון כי השיפור המדיד בביצועי המודל, המתבטא בעלייה של 1.11% במדד ה-F1-score (מ-0.9651 ל-0.9762), עשוי לנבוע במידה מסוימת מ overfitting-, קיימים מספר גורמים המצביעים על תקפות ורלוונטיות השיפור.

ראשית, העובדה כי השיפור בביצועים הינו מתון ועקבי מצביעה על כך שהמודל לא "למד יתר על המידה" את מאפייני הדאטה החדש. שנית, וחשוב מכך, הנתונים ששימשו לאימון הנוסף מייצגים דפוסי תקשורת עדכניים ומקומיים, מה שמאפשר למודל להתאים את עצמו לאתגרי אבטחת המידע העכשוויים. התאמה זו חיונית במיוחד בתחום זיהוי הפישינג, שבו שיטות ההתקפה מתפתחות ומשתנות באופן תדיר.

יתרה מזאת, העובדה שהנתונים נלקחו מתוך הקשר אמיתי ורלוונטי מחזקת את תקפות המודל בסביבת העבודה המיועדת. בניגוד למודלים המתבססים אך ורק על מאגרי נתונים סינתטיים או היסטוריים, המודל המשופר משקף טוב יותר את המציאות היומיומית של ניסיונות פישינג עכשוויים, ובכך מספק הגנה מותאמת יותר לאיומים הנוכחיים.

חשוב לציין כי בעתיד, עם המשך השימוש במערכת והאימון המתמשך של מודל ה-DistilBERT על נתונים אמיתיים שיתקבלו מהמשתמשים, צפויה מגמת שיפור מתמדת בדיוק המודל. שיפור זה יתבסס על למידה מתמשכת מדפוסי פישינג עדכניים ואמיתיים, ולא יהיה תוצר של ,overfitting אלא של התאמה הולכת וגדלה למציאות המשתנה של איומי הסייבר. גישה זו של למידה מתמשכת תאפשר למודל להתפתח ולהשתפר באופן אורגני, תוך שמירה על רלוונטיות ואפקטיביות לאורך זמן.

# תוצאות והערכה

## מדדים להשוואת המודלים

בכדי להעריך את ביצועי המודלים שלנו בפרויקט ולבחור את המודל האפקטיבי ביותר לזיהוי הודעות פישינג, השתמשנו במספר מדדים מקובלים בתחום הלמידת מכונה. כל אחד מהמדדים הללו מציג היבט מסוים של ביצועי המודל, ומאפשר לנו להבין את יכולתו להתמודד עם המשימה באופן מקיף ומדויק.

1. **דיוק (Accuracy)** : מדד הדיוק מתאר את היחס בין מספר הדוגמאות שסווגו בצורה נכונה לבין סך כל הדוגמאות במערכת הנתונים. כלומר, הוא מתאר את אחוז ההודעות שהמודל הצליח לסווג בצורה נכונה, בין אם הן הודעות פישינג ובין אם הן הודעות לגיטימיות. למרות שמדד זה פשוט להבנה, הוא עלול להטעות במקרים של מאגרי נתונים לא מאוזנים, כמו במקרה שלנו, שבו כמות הודעות הפישינג קטנה בהרבה מכמות ההודעות הלגיטימיות.
2. **דיוק סיווג ההודעות הפישינג (Precision):** מדד ה-Precision מתאר את היחס בין מספר ההודעות שסווגו בצורה נכונה כהודעות פישינג לבין סך כל ההודעות שסווגו כהודעות פישינג על ידי המודל. כלומר, הוא מודד את דיוק הסיווג עבור קטגוריית הפישינג בלבד - כמה מההודעות שהמודל סיווג כפישינג באמת היו פישינג. מדד זה חשוב במיוחד כדי להקטין את מספר ההודעות הלגיטימיות שמסווגות בטעות כפישינג (False Positives).
3. **רגישות או True Positive Rate (Recall):** מדד ה-Recall מציג את היחס בין מספר ההודעות שסווגו נכון כהודעות פישינג לבין סך כל ההודעות שהיו פישינג בפועל במערכת הנתונים. במילים אחרות, הוא מציג את יכולת המודל לזהות את כל ההודעות הפישינג באמת. מדד זה קריטי במקרה שבו חשוב לנו לזהות כמה שיותר ניסיונות פישינג, גם במחיר של סיווג שגוי של הודעות לגיטימיות (False Negatives).
4. **F1-Score:** מדד ה-F1 משלב בין מדדי ה-Precision וה-Recall ומטרתו לספק תמונה מאוזנת יותר של ביצועי המודל כאשר יש חוסר איזון בין חשיבותם של שני המדדים. המדד מחושב כממוצע הרמוני של Precision ו-Recall והוא נותן משקל שווה לשניהם. ערך גבוה של F1 מצביע על כך שהמודל מצליח גם לזהות ניסיונות פישינג באופן (Recall גבוה) וגם להיות מדויק בסיווגים שלו (Precision גבוה) ולכן מהווה מדד חשוב להערכת ביצועים כוללת.

השימוש במדדים אלו מאפשר לנו להעריך את ביצועי המודלים מנקודות מבט שונות, ולבחור את המודל שמספק את האיזון הטוב ביותר בין זיהוי ניסיונות פישינג ובין הימנעות מסיווג שגוי של הודעות לגיטימיות כפישינג.

## תוצאות הניסויים

בפרויקט ביצענו ניסויים שונים על מנת להעריך את ביצועי המודלים שלנו במטרה לזהות הודעות פישינג. לצורך כך השווינו בין שתי גישות מרכזיות – גישת למידת מכונה קלאסית שכללה מספר אלגוריתמים כגון Logistic Regression, KNN, Random Forest ועוד, וגישת עיבוד שפה טבעית (NLP) באמצעות מודל טרנספורמטיבי מתקדם (DistilBERT).

**תוצאות הניסויים וההבדלים בין הגישות**

המודלים הקלאסיים של למידת מכונה הצליחו להשיג תוצאות טובות בזיהוי הודעות פישינג, כאשר המודל הטוב ביותר מבין המודלים הקלאסיים היה XGBoost (עם Random State 123). מודל זה השיג F1-score של 0.8823, דיוק של 0.9045, ו-Recall של 0.9168, מה שמצביע על יכולת גבוהה הן לזיהוי הודעות פישינג והן לדיוק בסיווגן. יש להדגיש כי XGBoost הוא מודל המסוגל להבחין בין תבניות מורכבות, במיוחד בסביבה של כמות נתונים גדולה, מה שעשוי להסביר את ביצועיו הטובים.

מודל ה-DistilBERT שנעשה בו שימוש במסגרת הגישה של עיבוד שפה טבעית, הראה שיפור בביצועים לאורך שלושת האפוקים של האימון. ה-F1-Score שהתקבל בסוף תהליך האימון עמד על 0.9762, מה שמציין שיפור ניכר בביצועים לעומת המודלים הקלאסיים. מודל DistilBERT הצליח להבין את השפה בצורה עמוקה יותר ולהבחין בהקשרים שונים בהודעות, מה שאפשר לו להגיע לביצועים טובים יותר במונחים של דיוק ורגישות.

ניתן לראות שהגישה של עיבוד שפה טבעית, ובפרט שימוש במודלים טרנספורמטיביים כמו DistilBERT, השיגה תוצאות טובות יותר בהשוואה ללמידת מכונה קלאסית. הסיבה לכך היא שמודלי הטרנספורמציה מסוגלים לנתח את הטקסט בצורה מתוחכמת יותר, להבנת ההקשרים הלשוניים השונים ולזיהוי מאפיינים ייחודיים המאפיינים הודעות פישינג.

עם זאת, יש לקחת בחשבון את עלות האימון של מודל כמו DistilBERT , הכוללת דרישות חישוביות גבוהות יותר וזמן ריצה ארוך יותר לעומת המודלים הקלאסיים. מאידך, המודלים הקלאסיים של למידת מכונה מהירים יותר לאימון, אך מוגבלים יותר ביכולתם להבחין בין ניואנסים בטקסט.

במהלך הניסויים, ראינו גם שהמודלים הפשוטים יותר, כמו Logistic Regression ו-Naive Bayes היו פחות מדויקים בזיהוי הודעות פישינג, דבר שיכול לנבוע מהמורכבות של הודעות הפישינג והשימוש בתחבולות טקסטואליות המאפיינות אותן.

**מסקנות**

המודל הטוב ביותר היה מודל DistilBERT במסגרת השימוש בגישת ה-NLP, שהצליח להשיג ביצועים גבוהים במיוחד בזיהוי הודעות פישינג עם F1-score של 0.9762. מודל XGBoost, במסגרת המודלים של למידת מכונה, גם הוא השיג תוצאות טובות עם F1-score של 0.8823, והוכיח שהוא מתאים כאשר נדרש איזון בין דיוק ובין זמן ריצה ותשתית חישובית.

תמונה שמכילה טקסט, צילום מסך, צבעוני, תרשים

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, קו, תרשים, עלילה

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך, תרשים, מלבן

התיאור נוצר באופן אוטומטי

## דיון בתוצאות

תוצאות הניסויים מצביע על כך ששימוש במודלים מבוססי עיבוד שפה טבעית דוגמת DistilBERT, מציע יתרון משמעותי בזיהוי הודעות פישינג בהשוואה למודלים קלאסיים של למידת מכונה. מודלים אלה הצליחו להבין את ההקשרים הלשוניים המורכבים והציגו דיוק ורגישות גבוהים יותר, כפי שנראה במדד F1 הגבוה שהתקבל. עם זאת, למודלים הקלאסיים יש יתרון של פשטות ומהירות באימון, אך הם מוגבלים ביכולת להבין את המורכבות של שפה טבעית, מה שמקשה עליהם לזהות הודעות פישינג מתוחכמות. לפיכך, למרות העלות החישובית הגבוהה, השימוש ב-NLP מהווה פתרון מבטיח ואפקטיבי יותר להגנה על משתמשים מפני הונאות פישינג.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Random State | Accuracy | Precision | Recall | F1 Score |
| Naive Bayes | 42 | 63.80% | 65.08% | 15.75% | 25.36% |
| Naive Bayes | 123 | 65.05% | 70.50% | 18.05% | 28.75% |
| Logistic Regression | 42 | 77.95% | 67.93% | 82.46% | 74.49% |
| Logistic Regression | 123 | 79.00% | 68.70% | 84.89% | 75.95% |
| KNN | 42 | 81.45% | 75.56% | 77.59% | 76.56% |
| KNN | 123 | 82.55% | 76.60% | 79.64% | 78.09% |
| Decision Tree | 42 | 83.80% | 75.53% | 86.56% | 80.67% |
| Decision Tree | 123 | 84.50% | 77.04% | 85.92% | 81.23% |
| Gradient Boosting | 42 | 88.90% | 83.16% | 89.76% | 86.33% |
| Neural Network | 42 | 88.90% | 83.16% | 89.76% | 86.33% |
| Neural Network | 123 | 88.95% | 85.26% | 86.68% | 85.97% |
| XGBoost | 42 | 89.10% | 83.00% | 90.65% | 86.66% |
| Random Forest | 42 | 89.55% | 84.05% | 90.40% | 87.11% |
| Gradient Boosting | 123 | 89.70% | 83.70% | 91.42% | 87.39% |
| Random Forest | 123 | 89.80% | 84.63% | 90.27% | 87.36% |
| XGBoost | 123 | 90.45% | 85.04% | 91.68% | 88.23% |
| DistilBERT (NLP) | **Epoch 3** | 98.00% | 96.00% | 99.00% | 97.62% |

# ממשק UI למשתמש קצה

בהמשך הפרויקט פיתחנו ממשק משתמש (UI) המאפשר למשתמשים קצה לבצע ניתוחים על הודעות דוא"ל ולבחון האם הן מהוות ניסיונות פישינג. ממשק ה-UI נבנה באמצעות ספריית Tkinter בשפת Python, תוך מתן דגש על נוחות, אינטואיטיביות ושימושיות גבוהה.

**תכונות ויכולות הממשק**

הממשק כולל מגוון פיצ'רים המאפשרים חווית שימוש מלאה ואינטראקטיבית:

1. **טעינת ושמירת הודעות**: הממשק מאפשר למשתמשים לטעון קבצי טקסט מקומיים המכילים את תוכן הדוא"ל לניתוח. בנוסף, ניתן לשמור את התוכן לאחר עריכה או ניתוח כקובץ טקסט להמשך עבודה.
2. **ניתוח הודעה**: המשתמש מזין את טקסט ההודעה לתוך אזור קלט ייעודי, ולוחץ על כפתור "Analyse Email" לצורך ניתוח ההודעה. לאחר הלחיצה, המערכת מבצעת חיזוי על ידי מודל מבוסס DistilBERT ומספקת את ההסתברויות לכך שההודעה היא פישינג או בטוחה.
3. **חישוב הסתברויות ותצוגת תוצאות**: המערכת מציגה את ההסתברויות לכך שההודעה היא פישינג או בטוחה בצורה ברורה ובולטת. בנוסף, הממשק נותן אינדיקציה לרמת הסיכון עם צבעים מתאימים (ירוק עד אדום) והמלצות מתאימות לפעולה בהתאם לרמת הסיכון.
4. **ניהול היסטוריה**: הממשק כולל אופציה להצגת היסטוריית הניתוחים שבוצעו, כך שהמשתמשים יכולים לצפות בניתוחים קודמים כולל התאריך והשעה שבהם בוצעו, ההסתברויות המחושבות, ותצוגה חלקית של תוכן ההודעה. היסטוריית הניתוחים נשמרת בקובץ JSON.
5. **עזרה למשתמש**: קיימת אפשרות לגשת לחלון "עזרה", שבו מוסברים שלבי השימוש במערכת בצורה פשוטה וברורה, כמו גם מידע על הגרסה של המערכת והמפתחים שלה.

**תיאור פעולת הממשק**

לאחר הזנת ההודעה, המערכת מבצעת חיזוי באמצעות המודל DistilBERT אשר אומן לזהות מאפיינים טיפוסיים של הודעות פישינג. בהתאם לחיזוי, מוצגות ההסתברויות שההודעה היא פישינג או בטוחה, ולאחר מכן ניתנת המלצה למשתמש כיצד לפעול. בנוסף, כל ניתוח נשמר בהיסטוריית הניתוחים, מה שמאפשר גישה למידע הנוגע להודעות שנבדקו בעבר.

ממשק המשתמש נועד להקל על משתמשי קצה להתמודד עם הסיכונים הקשורים להודעות פישינג, באמצעות מתן פידבק מהיר ואפקטיבי על הסיכון הכרוך בהודעה שהתקבלה. היעד המרכזי היה לעזור למשתמשים לקבל פידבק ברור ומדויק על הודעות דוא"ל, תוך מינימום זמן ומאמץ. הממשק פותח בצורה שתאפשר למשתמשים ללא ידע טכני לקבל תובנות במהירות ובאופן ידידותי, ובכך לסייע להם בקבלת החלטות מושכלות לגבי הפעולות שעליהם לבצע בהקשר להודעה.

הממשק מתאפיין בעיצוב נקי וקל להבנה, עם דגש על נוחות השימוש. כל הפעולות המרכזיות – טעינה, שמירה, ניתוח והצגת התוצאות – מרוכזות בצורה נגישה וברורה על גבי המסך, תוך שמירה על סידור לוגי ואינטואיטיבי של הפיצ'רים השונים. המשתמשים אינם נדרשים להתמודד עם פרטים טכניים מורכבים, והכלים לניתוח הודעות נגישים בלחיצת כפתור אחת בלבד.

העיצוב של הממשק מכיל אלמנטים ויזואליים שמסייעים בהעברת המידע בצורה ברורה וויזואלית, כמו שימוש בצבעים שמייצגים את רמות הסיכון והצגת המלצות פעולה מותאמות על פי רמת הסיכון. כל אלו מסייעים להפוך את תהליך קבלת הפידבק לפשוט, מובן ובזמן אמת, כך שהמשתמשים יכולים להבין מיד האם מדובר בסיכון גבוה ומה עליהם לעשות.

לממשק חוויית שימוש זורמת, פשוטה ומקיפה, שמאפשרת למשתמשים להרגיש בטוחים יותר בעבודה מול הודעות חשודות, ולקבל החלטות נכונות תוך שימוש במידע מהימן ומבוסס ניתוחים מתקדמים של המודל.

תמונה שמכילה טקסט, חשמל, צילום מסך, תוכנה

התיאור נוצר באופן אוטומטי



## פונקציונאליות נוספת למשתמש הקצה

במסגרת מערכת PhishGuard AI, קיימת היכולת לניתוח מקבילי של מספר הודעות דואר אלקטרוני. המערכת תומכת בטעינה וניתוח של מספר הודעות בו-זמנית, כולל קבצי MSG, מה שמייעל משמעותית את תהליך העבודה עבור המשתמש.

הממשק מציג למשתמש טבלה מרוכזת של ההודעות שנותחו, הכוללת את דירוג הסיכון וההסתברות לפישינג עבור כל הודעה, יחד עם המלצות רלוונטיות. תכונה זו שימושית במיוחד עבור משתמשים המקבלים מספר רב של הודעות דואר אלקטרוני ומעוניינים לבצע סריקה מהירה ויעילה של מספר הודעות חשודות.

היסטוריית הניתוחים נשמרת באופן מאורגן, המאפשר למשתמשים לעקוב אחר הודעות שנבדקו בעבר ולזהות דפוסים חוזרים. פונקציונליות זו הופכת את המערכת לנוחה יותר לשימוש ומותאמת לצרכי המשתמשים בסביבת העבודה היומיומית.